Birth and Death of Domains: Simple Models of Genome Evolution Explain Power Law Distribution of Protein Family Size

Birth, Death and Innovation Models (BDIM) of genome evolution

Eugene V. Koonin
NCBI, NLM, NIH, Bethesda, MD

\[
df_i(t)/dt = +\lambda f_{i-1} - \delta f_i - \delta f_{i+1} + \lambda f_{i+1}
\]

- mathematical formulation and classification of models
- where do the power laws come from
- empirical data fit
- do BDIM tell us anything biologically non-trivial?
BDIM: basic concepts

part-of-gene-encoding-an-individual-domain = "domain"

gene

Genome: a finite "bag" of independently evolving domains.

Domains form families of paralogs (domain families for short).

BDIM: elementary events

Birth

Death

Innovation

BDIM – Birth, Death and Innovation Model
Birth and Death of Domains: Simple Models of Genome Evolution Explain Power Law Distribution of Protein Family Size

**BDIM: the layout of the model**

- **Domain innovation rate**: λ
- **Domain family birth rate**: λ
- **Per-family death rate**: δ
- **Number of families in a size class**: f
- **Maximum number of domains in a family**: N

**BDIM: the basic equations**

- **Rate of change**: df(t)/dt

  - Gain: birth in class i-1
    \[
    df_{i-1}(t)/dt = λ_{i-1}f_{i-1} - λ_{i}f_i + δ_{f_{i}}
    \]

  - Loss: birth in class i
    \[
    df_i(t)/dt = λ_i f_i - δ_i f_i + δ_{f_{i+1}} f_{i+1}
    \]

  - ...  

  - Gain: death in class i+1
    \[
    df_{i+1}(t)/dt = λ_{i+1} f_{i+1} - δ_{f_{i+1}}
    \]

- **No birth into and death from class N+1**

- **Loss: death in class i**

- **Total number of families**: F(t) = ∑ f_i(t)

Dr. Eugene Koonin, NIH (KITP Bio Networks 2/25/03)
BDIM: the equilibrium

\[
df_i(t)/dt = 0 \quad \text{equilibrium for the number of domain families in each size class}
\]
\[
dF(t)/dt = 0 \quad \text{equilibrium for the total number of families}
\]

• There exists a unique and stable equilibrium state \( f_1, f_2, \ldots, f_N \)
• The model reaches equilibrium exponentially:
\[
|f_i(t) - f_i| \sim e^{-\kappa t}
\]
• The model is "open" at one end only (class 1 families).

A simple condition describes the equilibrium for the total number of families:

\[
\nu = \delta f_1
\]

innovation

Death of \textbf{families}

Simple BDIM

Independence hypothesis:

i. all elementary events are independent of each other
ii. the rates of \textit{individual domain} birth (\( \lambda \)) and death (\( \delta \)) do not depend on \( i \) (number of domains in a family).

Corollary:

\[
\lambda_i = \lambda i
\]
\[
\delta_i = \delta i
\]

The basic equation for domain family evolution:

\[
df_i(t)/dt = \lambda (i-1)f_{i-1} - (\delta + \lambda)if_i + \delta (i+1)f_{i+1}
\]
Simple BDIM: equilibrium

The equilibrium solution:

\[ f_i \sim (\lambda/\delta)/i \] - truncated logarithmic distribution

If \( \lambda = \delta \)

\[ f_i \sim 1/i \] - power law (degree = -1)

BDIM hierarchy

- Master BDIM: \( \lambda_i = \lambda(i) \)
- Rational: \( \delta_i = \delta(i) \)
- Polynomial: \( \lambda_i = P(i)/Q(i) \)
- Simple: \( \lambda_i = \lambda i \)
- Linear: \( \lambda_i = \lambda(i+a) \)
- Quadratic: \( \lambda_i = \lambda(i+a)(i+a_1) \)
BDIM: balanced and non-balanced models

Master BDIM
\[ \chi(i) = \lambda_i i^{\delta_i} = i^\theta (1 + a/i + O(1/i^2)) \]

- \( s \neq 0 \)
  - non-balanced
  - \( f_i \sim i^{\delta_i} \theta_i^{i^\alpha} \)

- \( s = 0 \)
  - \( \theta \neq 1 \)
    - first-order balanced
    - \( f_i \sim i^{\delta_i} \theta_i^{i^\alpha} \)
  
  - \( \theta = 1 \)
    - \( a \neq 0 \)
      - second-order balanced
      - \( f_i \sim i^\alpha \)
    
    - \( a = 0 \)
      - high-order balanced
      - \( f_i \sim \text{const} \)

BDIM: only first/second order balanced models make sense

- Non-balanced BDIM: unrealistic family size distributions with extremely strong dependence on \( i \):
  - either no large families at all or mostly large families
- High-order balanced BDIM: equally unrealistic, uniform distribution
Birth and Death of Domains: Simple Models of Genome Evolution Explain Power Law Distribution of Protein Family Size

**Linear BDIM**

\[ \lambda_i = \lambda(i+a) \]
\[ \delta_i = \delta(i+b) \]
\[ \theta = \frac{\lambda}{\delta} \]

\[ f = \frac{\Gamma(1+b)}{\Gamma(1+a)} \left( \frac{\nu}{\delta} \right)^{\theta-1} \frac{\Gamma(i+a)}{\Gamma(i+b+1)} \]

constant exponent \( \sim i^{a-b-1} \) for large \( i \)

\[ \theta = 1 \text{ (power law asymptote, } k = a-b-1) \]

\[ \theta < 1, a = b \text{ (truncated logarithmic)} \]

**Empirical Data: Domains in Genomes**

- CDD PSSMs
- E. coli
- H. sapiens
- A. thaliana

Domain family size distribution

E < 0.001

RPS-BLAST hits

Dr. Eugene Koonin, NIH (KITP Bio Networks 2/25/03)
Birth and Death of Domains: Simple Models of Genome Evolution Explain Power Law Distribution of Protein Family Size

Data Fitting

- strong deviation from a straight line
- data "hitting the bottom" near the tail

Data Fitting: Choosing the Model of Evolution

*E. coli*
- Linear BDIM: $P(\chi^2) = 0.71$
- Simple BDIM: $P(\chi^2) = 3 \times 10^{-6}$
- Power law: $P(\chi^2) = 2 \times 10^{-10}$
Data Fitting: Choosing the Model of Evolution

A. thaliana
Linear BDIM: $P(\chi^2) = 0.09$
Simple BDIM: $P(\chi^2) = 6 \times 10^{-13}$
Power law: $P(\chi^2) = 2 \times 10^{-54}$

Data Fitting:
Power Approximation vs Power Asymptote under the linear BDIM
Birth and Death of Domains: Simple Models of Genome Evolution Explain Power Law Distribution of Protein Family Size

Linear BDIM parameters and other relevant numbers for bacteria, archaea and eukaryotes

<table>
<thead>
<tr>
<th></th>
<th>noRF</th>
<th>nFam</th>
<th>nDom</th>
<th>nHit</th>
<th>Max</th>
<th>d_i (d'_i)</th>
<th>a</th>
<th>b</th>
<th>k</th>
<th>v/δ ν/λ</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sc</td>
<td>6340</td>
<td>1080</td>
<td>4575</td>
<td>3331</td>
<td>130</td>
<td>420 (436)</td>
<td>1.55</td>
<td>3.27</td>
<td>-2.72</td>
<td>1861.8</td>
<td>[3.28..3.53]</td>
</tr>
<tr>
<td>Dm</td>
<td>13605</td>
<td>1405</td>
<td>11734</td>
<td>7262</td>
<td>335</td>
<td>426 (435)</td>
<td>1.62</td>
<td>2.79</td>
<td>-2.17</td>
<td>1648.2</td>
<td>[8.44..15.50]</td>
</tr>
<tr>
<td>Ce</td>
<td>20524</td>
<td>1418</td>
<td>17054</td>
<td>11090</td>
<td>662</td>
<td>423 (421)</td>
<td>1.13</td>
<td>2.03</td>
<td>-1.89</td>
<td>1273.0</td>
<td>[16.18..∞]</td>
</tr>
<tr>
<td>At</td>
<td>25854</td>
<td>1405</td>
<td>21238</td>
<td>15006</td>
<td>1535</td>
<td>270 (277)</td>
<td>3.80</td>
<td>4.98</td>
<td>-2.18</td>
<td>1657.7</td>
<td>[17.09..26.89]</td>
</tr>
<tr>
<td>Hs</td>
<td>39883</td>
<td>1681</td>
<td>27844</td>
<td>16755</td>
<td>1151</td>
<td>298 (288)</td>
<td>5.16</td>
<td>6.43</td>
<td>-2.27</td>
<td>2136.2</td>
<td>[17.14..22.88]</td>
</tr>
<tr>
<td>Thema</td>
<td>1846</td>
<td>772</td>
<td>1683</td>
<td>1268</td>
<td>97</td>
<td>501 (499)</td>
<td>0.14</td>
<td>2.22</td>
<td>-3.08</td>
<td>1606.4</td>
<td>[1.04..1.06]</td>
</tr>
<tr>
<td>Meth</td>
<td>1869</td>
<td>693</td>
<td>1480</td>
<td>1150</td>
<td>43</td>
<td>438 (436)</td>
<td>0.12</td>
<td>2.00</td>
<td>-2.88</td>
<td>1305.3</td>
<td>[1.18..1.27]</td>
</tr>
<tr>
<td>Sulso</td>
<td>2977</td>
<td>695</td>
<td>1950</td>
<td>1614</td>
<td>81</td>
<td>386 (385)</td>
<td>0.36</td>
<td>2.04</td>
<td>-2.68</td>
<td>1167.8</td>
<td>[1.83..2.00]</td>
</tr>
<tr>
<td>Bacsu</td>
<td>4100</td>
<td>1002</td>
<td>3413</td>
<td>2502</td>
<td>124</td>
<td>507 (510)</td>
<td>0.48</td>
<td>2.01</td>
<td>-2.53</td>
<td>1534.6</td>
<td>[2.46..2.79]</td>
</tr>
<tr>
<td>Escco</td>
<td>4289</td>
<td>1078</td>
<td>3624</td>
<td>2765</td>
<td>140</td>
<td>523 (519)</td>
<td>0.84</td>
<td>2.54</td>
<td>-2.70</td>
<td>1837.0</td>
<td>[2.45..2.61]</td>
</tr>
</tbody>
</table>

G = total duplication rate/innovation rate

Linear BDIM: Size Does Matter?

\[
\frac{\lambda}{i} = \frac{\lambda(1+a_i)}{i} \text{ per domain birth rate}
\]

\[
\frac{\delta}{i} = \frac{\delta(1+b_i)}{i} \text{ per domain death rate}
\]
Conclusions

I. Only balanced BDIM produce reasonable equilibrium distributions of domain family size; equilibrium is reached rapidly, suggesting a "punctuated equilibrium"-like mode of genome evolution.

II. The simplest evolutionary model that adequately describes the observed distribution of domain family size is the linear, second-order balanced BDIM; accordingly, per-domain birth/death rate depends on family size, the larger families being less dynamic in evolution.

III. The rates of domain innovation and birth are comparable.

The original version of BDIM is fully deterministic.

In order to be able to explore the dynamics of genome evolution, we introduce a stochastic (Markov) version.
Birth and Death of Domains: Simple Models of Genome Evolution Explain Power Law Distribution of Protein Family Size

Markov version of BDIM (0 class introduced)
(innovation interpreted as extraction from class 0)

\[
\begin{align*}
\frac{dp_0(t)}{dt} &= -\lambda_0 p_0(t) + \delta_1 p_1(t), \\
\frac{dp_1(t)}{dt} &= \lambda_0 p_0(t) - (\lambda_1 + \delta_1) p_1(t) + \delta_2 p_2(t), \\
\vdots \\
\frac{dp_i(t)}{dt} &= \lambda_i p_{i-1}(t) - (\lambda_i + \delta_i) p_i(t) + \delta_{i+1} p_{i+1}(t) \quad \text{for } 1 < i < N, \\
\frac{dp_N(t)}{dt} &= \lambda_{N-1} p_{N-1}(t) - \delta_N p_N(t)
\end{align*}
\]

Modified Markov version of BDIM (no 0 class, class 1 immortal)

\[
\begin{align*}
\frac{dp_1(t)}{dt} &= -\lambda_1 p_1(t) + \delta_2 p_2(t), \\
\frac{dp_i(t)}{dt} &= \lambda_i p_{i-1}(t) - (\lambda_i + \delta_i) p_i(t) + \delta_{i+1} p_{i+1}(t) \quad \text{for } 1 < i < N, \\
\frac{dp_N(t)}{dt} &= \lambda_{N-1} p_{N-1}(t) - \delta_N p_N(t).
\end{align*}
\]

no innovation

Probability for a family to reach size \( n \) from size \( i \)
before extinction (size 0)

\[
P(i; n) = \frac{1}{(1 + \sum_{j=1}^{i-1} \prod_{k=1}^{j} \delta_k/\lambda_k)} \left(1 + \sum_{j=0}^{n-1} \prod_{k=1}^{j} \delta_k/\lambda_k \right)
\]

And, for 2nd order balanced linear BDIM and \( i=1 \),

\[
P(1, n) = \frac{1}{(1 + (1 + a) \left( \frac{\Gamma(1 + a)}{\Gamma(1 + b)} \cdot \frac{\Gamma(b + n + 1)}{\Gamma(a + n)} \cdot \frac{\Gamma(2 + b)}{\Gamma(1 + a)} \right)}
\]

Dr. Eugene Koonin, NIH (KITP Bio Networks 2/25/03)}
Birth and Death of Domains: Simple Models of Genome Evolution Explain Power Law Distribution of Protein Family Size

#### Small families

- **Probability that a singleton forms a family with $i$ members via a series of duplications before dying (double-log scale) under 2nd order balanced linear BDIM**

#### Full range of families

- **Probability that a singleton forms a family with $N$ members via a series of duplications before dying (double-log scale) under 2nd order balanced linear BDIM**
Birth and Death of Domains: Simple Models of Genome Evolution Explain Power Law Distribution of Protein Family Size

**Time before extinction of a family of size \( n \) for 2\textsuperscript{nd} order balanced linear BDIM**

\[
E(n) = \frac{1}{\lambda} \sum_{j=1}^{n} \frac{\Gamma(j+b)}{\Gamma(j+a)} \sum_{i=j}^{n} \frac{\Gamma(k+a)}{\Gamma(k+1+b)}
\]

For the time being, we use \( 1/\lambda \) as a natural time scale for BDIM…

**Mean time of extinction \( E(n) \) depending on family size for 2\textsuperscript{nd} order balanced linear BDIM**
Birth and Death of Domains: Simple Models of Genome Evolution Explain Power Law Distribution of Protein Family Size

Time required for a singleton to reach family size \( n \) under 2nd order balanced linear BDIM (measured in \( 1/\lambda \) units)

\[
M(1;n) = \frac{1}{\lambda} \sum_{k=1}^{n-1} \left( \frac{\Gamma(b+k+1)}{\Gamma(a+k+1)} \right) \sum_{i=1}^{k} \frac{\Gamma(a+i)}{\Gamma(b+i+1)}.
\] (8.7)

Mean time for the formation of families depending on \( n \) under the 2nd order balanced linear BDIM
Birth and Death of Domains: Simple Models of Genome Evolution Explain Power Law Distribution of Protein Family Size


The Evolutionary Fate and Consequences of Duplicate Genes

0.002 to 0.02 duplications per gene per million years

Mean time for the formation of families depending on $n$ under the 2nd order balanced linear BDIM

Assuming duplication rate $2 \times 10^{-8}$/yr,

$T(1,N) \sim 10^{13}$ yr, i.e., $\sim 1,000$ x age of Universe
Stochastic characteristics of the linear BDIM:
i) extremely large difference between times of formation and
extinction of the largest families for some genomes –
extinction happens much faster;
ii) Under the available empirical estimates of duplication rate,
the time required for the formation of the largest families is
unrealistically long

Consequently, we must replace the linear BDIM with another model
such that:
1) the stationary distribution of the family sizes is the same as for
the linear BDIM;
2) the new model provides for much more rapid evolution of
gene families under realistic values of duplication and deletion rates;
3) the ratio of family formation and extinction mean times must be
significantly less than for the linear BDIM.
To obtain a new BDIM without changing the stationary distribution:

\[ \lambda^*_{i} \rightarrow \lambda_{i} g(i), \quad \delta^*_{i} \rightarrow \delta_{i} g(i-1) \]

\[ g > 0 \]
\[ g(0) = 1 \]

Probability of formation of family of size \( n \) prior to extinction:

\[ P^*(1,n) = 1 / \left( 1 + \frac{\Gamma(1+a)}{\Gamma(1+b)} \sum_{i=1}^{n-2} \frac{1}{g(k)} \frac{\Gamma(b+k+1)}{\Gamma(a+k+1)} \right) \]

Extinction time:

\[ E^*_s = \frac{1}{\lambda} \sum_{k=a}^{n} \sum_{i=k}^{N} \frac{1}{g(k-1)} \left[ \frac{\Gamma(a+i)}{\Gamma(b+i+1)} \frac{\Gamma(b+k)}{\Gamma(a+k)} \right] \]

Family formation time (class 1 immortal):

\[ M^*(1,n) = \frac{1}{\lambda} \sum_{i=a}^{n} \frac{1}{g(k)} \frac{\Gamma(a+i)}{\Gamma(b+i+1)} \sum_{k=a}^{i} \frac{\Gamma(b+k)}{\Gamma(a+k+1)} \]

**Quadratic BDIM**

Let us consider the 2nd order balanced polynomial (quadratic) Markov BDIM with duplication and deletion rates

\[ \lambda_{i} = \lambda(i+a)(i+1), \]
\[ \delta_{i} = \lambda(i+b)i. \]

This may be interpreted as introducing pairwise interactions between family members.
Birth and Death of Domains: Simple Models of Genome Evolution Explain Power Law Distribution of Protein Family Size

Probability that a singleton forms a family with \( i \) members via a series of duplications before dying under the quadratic BDIM.

Mean time of extinction under the quadratic BDIM, \( E_2(n) \), depending on family size \( i \).

Dr. Eugene Koonin, NIH (KITP Bio Networks 2/25/03)
Birth and Death of Domains: Simple Models of Genome Evolution Explain Power Law Distribution of Protein Family Size

Assuming duplication rate ($\lambda$) $\sim 2 \times 10^{-8}$/yr, $T(1,N) \sim 10^{10}$ yr – about the age of Universe, still longer than real evolution times but not by much.

Probability $P_{d+1}(i,N)$ of formation of the largest family before extinction depending on the model degree $k=1+d$ and the initial family size; the values of parameters $a=1.55$, $b=2.79$, $N=335$ are for $Dme$. 
The dependence of $\ln(E_{d+1}(n))$ – extinction time - on family size, $1<n<335$, and on the model degree, $0<d<5$; the values of parameters $a=1.64$, $b=2.79$, $N=335$ are for $Dme$.

Mean time of family formation, $M_{d+1}(n)$ (log scale) for BDIM of different degrees (from linear to cubic): $d=0, 0.2, 0.5, 0.8, 1., 1.3, 1.5, 1.7, 2$ (from top to bottom); the parameter values are for $Dme$.
Birth and Death of Domains: Simple Models of Genome Evolution Explain Power Law Distribution of Protein Family Size

Mean time of family formation $M_{d+1}(n)$ depending on initial of family size ($1<n<335$) and model degree $d$; the parameter values $a=1.62$, $b=2.79$, $N=335$ are for $Dme$

Linear BDIM: $\lambda_i/i = \lambda(i+a)/i$, $\delta_i/i = \lambda(i+b)/i$.

Quadratic BDIM: $\lambda_i/i = \lambda(i+a)(i+1)/i$, $\delta_i/i = \lambda(i+b)$.
Birth and Death of Domains: Simple Models of Genome Evolution Explain Power Law Distribution of Protein Family Size

\[ \frac{\delta}{\lambda} \text{ ratio} \]

- Linear BDIM
- Quadratic BDIM
- Dependence on BDIM order and family size

Dr. Eugene Koonin, NIH (KITP Bio Networks 2/25/03)
Conclusions on stochastic (Markov) BDIMs and some general conclusions

2nd order balanced linear BDIM is sufficient to explain the observed distributions of domain family size.

However, when a stochastic model is used to estimate the time required to reach the maximum family size, the estimate is ~10,000 times greater than the time suggested by empirical data.

BDIMs of the degree 2-3 (quadratic/cubic) formally solve the problem by predicting evolutionary rate compatible with observations.

Conclusions on stochastic (Markov) BDIMs and some general conclusions

Biological interpretation of “interaction” between paralogs, which is intrinsic in higher-order BDIMs – does it reflect selection?

The dependence of birth and death rates on family size dramatically changes depending on BDIM order. This needs to be tested against detailed empirical analysis of paralogous families.
Some future directions…

- Incorporating selection into BDIM
- Combining BDIM with phylogenetic tree analysis
- Modeling evolutionary processes that lead to similar distributions in other contexts, e.g., multidomain protein architectures and interaction networks

“There are two kinds of science: physics and stamps collection”

Attributed to Ernest Rutherford
Birth and Death of Domains: Simple Models of Genome Evolution Explain Power Law Distribution of Protein Family Size

Acknowledgments

Georgy Karev (NCBI, NIH), Yuri Wolf (NCBI, NIH), Andrey Rzhetsky (Columbia University), Faina Berezovskaya (Howard University)